Review of NSF OCI EAGER and NSF OCI SDCI projects

M. Veeraraghavan
University of Virginia
mvee@virginia.edu
Sept. 30, 2011

This work was carried out as part of sponsored research projects from the NSF: OCI-1038058 and OCI-1127340

Projects

- NSF OCI EAGER: Towards increasing the usage of new high-speed network services by the scientific community
  - Participants:
    - UVA: Jie Li, Matt Manley, M. Veeraraghavan
    - NCAR: Steve Emmerson
- NSF OCI SDCI Net: An integrated study of datacenter and wide-area networking for distributed scientific computing
  - Participants:
    - UVA: Gradon Koelling, Zhenyang Liu, Peter Sahajian, M. Veeraraghavan
    - University of New Hampshire: Robert D. Russell
    - NCAR: John M. Dennis
- Program Manager: Kevin Thompson

Project Web sites

- EAGER:
  - Project web site: http://www.ece.virginia.edu/mv/research/EAGER/index.html
  - UVA Collab page:
    - https://collab.itc.virginia.edu/portal/site/2337694d-c6ec-4069-9e4d-1137758e3257
- SDCI
  - Project web site: http://www.ece.virginia.edu/mv/research/SDCI/index.html
  - UVA Collab page:
    - https://collab.itc.virginia.edu/portal/site/c23cb17a-1d23-4bbb-b9de-1e3a2b4ae505

Agenda

- **EAGER Project Review**
  - 9:30: EAGER introduction (MV)
  - 9:40: EAGER: IDD data characteristics (Matt Manley)
  - 9:55: EAGER: Year 1 report (MV)
  - 10:15: EAGER: Multicast transport protocol for VCs (Jie Li)
- **SDCI Project Review**
  - 10:30: SDCI year 1 workplan for UVA and UNH (MV)
  - 11:00: SDCI year 1 UCAR workplan: John Dennis, NCAR
  - 11:30: GridFTP, Ganglia, ANI testbed: Zhengyang Liu
  - 11:45: GridFTP data analysis: J. Gradon Koelling and Peter Sahajian
- 12-1: Lunch with discussion

Background

- UCAR Internet Data Distribution (IDD) project
  - Distributes real-time meteorology data
  - 10 GB/hour
  - Subscriber base: 170 institutions
  - Test case: bridging the gap between new network services and scientific applications
- Software used for distribution
  - Local Data Manager (LDM)
  - RPC based software
  - Uses unicast TCP connections from UCAR servers to each of the subscribing institutions, though a feed tree structure is supported

5

Data feedtypes

- Many feedtypes
  - http://www.unidata.ucar.edu/software/ldm/ldm-current/basics/feedtypes/
- Of these:
  - CONDUIT: NCEP high-resolution model output
  - GEM: Canadian Meteorological Center GEM model output
  - NEXRAD2: NEXRAD Level-II radar data

6

Motivation

Two Sample LDM Flows from Internet2 Netflow Data Analysis

Source IP	Destination IP	Source Port	Destination Port	IP Protocol	Bytes	Flow Duration
158.136.64.0	158.83.0.0	388	55842	6	23141060	2.2 Hours
128.117.136.0	129.186.184.0	388	43109	6	49995304	3.9 Hours

- Long duration of these flows made them seem well suited for dynamic circuit service (DCS)
- In current ESnet deployment of DCS, circuit setup delays are on the order of minutes

7

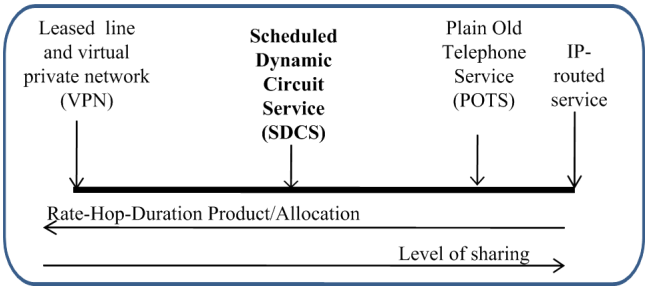
Problem statement

- Analyze the characteristics of the data distributed by IDD and determine the most suitable networking service to use for this data
- Modify LDM to use this service
- Undertake a macro-level cost/benefit analysis on the feasibility of adoption

8

Network services

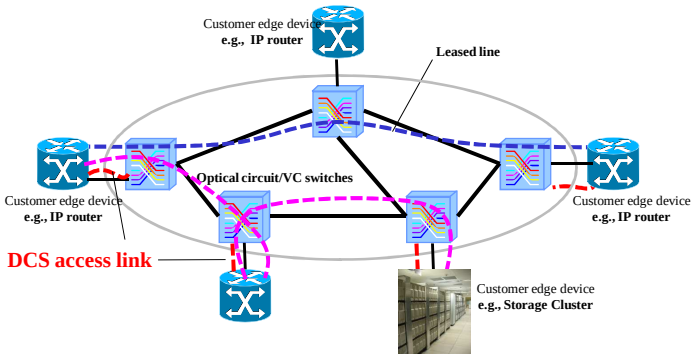
IEEE Comm. Magazine, Nov. 2010 paper, Veeraraghavan, Karol and Clapp (Telcordia)



- POTS: bufferless queueing system (M/G/m/m)
  - No duration specification in call setup
- SDCS: has a reservation scheduler
  - Need to specify call duration in request
  - Earliest Start Time and User Specified Start Times
  - Advance reservations a bonus!

9

Difference between leased-line and SDCS



- Leased-line: One contract: endpoints, rate, duration
- SDCS: Two contracts:
  - (1) long-term DCS access link (≡ IP access link)
  - (2) short-term VC to another DCS customer

10

Question

- Which of these four network services is best suited for IDD data?
  - IP routed service (+ TCP: reliable)
  - Static circuits (leased lines)
    - if continuous data flow, is this an option?
  - Scheduled dynamic circuit service (DCS)
    - if data flow is long-lived, option?
  - POTS: unscheduled DCS

11

To answer this question

- Per-flow data characteristic insufficient
  - typical classification:
    - loss-sensitive, high throughput
    - delay-sensitive, low latency
- Instead, need distribution topology
  - consider whole network view

12

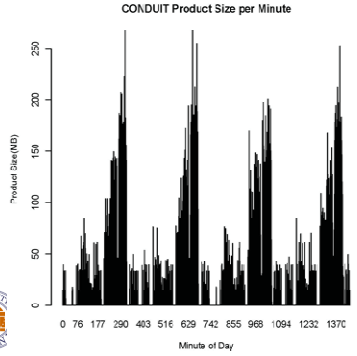
Real-time statistics

- http://www.unidata.ucar.edu/software/idd/rtstats/
- Ordered set of feedtypes (volume)
  - http://www.unidata.ucar.edu/cgi-bin/rtstats/rtstats_summary_volume?oliver.unidata.ucar.edu

13

CONDUIT data

- Installed and configured the LDM to receive CONDUIT data from UCAR: Jie Li
- Parsed and analyzed the log files for received data(9 sample days)
  - Peak throughput: 250 MB/minute (SD: 28.8 MB/minute)



14

Distribution structure

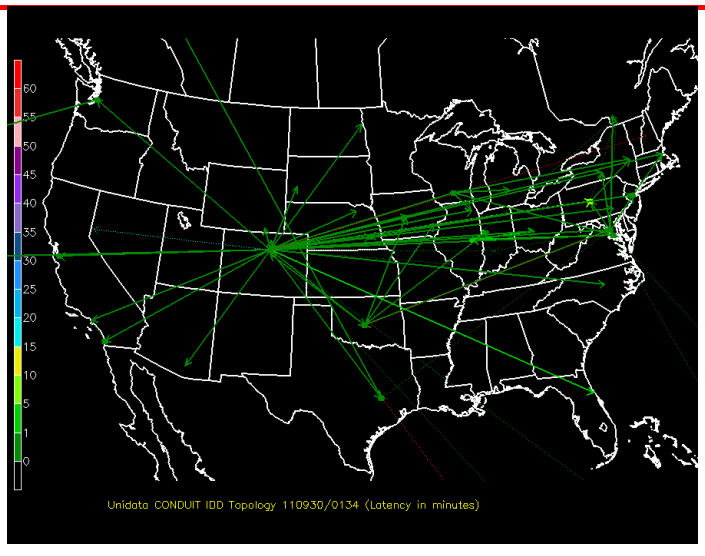
- Downloaded and parsed real-time statistics of the CONDUIT feed tree
- Data Distribution Topology of the CONDUIT feedtype
  - For the max fan-out of 104 receivers, the peak bandwidth requirement is 104 X 250 MB/minute ≈ 3.5Gbps
  - This is just for a single feedtype of a single application

CONDUIT Feed Tree Topology Information

Parameter	Number
# Distinct Hosts	163
# Sender Hosts	57
# Receiver Hosts	141
Max. Fan-out Number	104*

This maximum fan-out number comes from the UCAR site (idd.unidata.ucar.edu)

CONDUIT topology http://www.unidata.ucar.edu/cgi-bin/rtstats/rtstats_topogif?CONDUIT



16

Answer to question

- Different network service types
  - Static unicast VCs: unsuitable
    - Divide NCAR access link bandwidth between 104 subscribers: if 10 Gbps, then ~10 Mbps per subscriber
    - Subscribers would like to receive the data asap (low rate VC will increase latency)
  - Dynamic unicast VCs: unsuitable
    - For the worst-case fanout of 104, the total delay will be greater than with IP service, since for each receiver a new circuit needs to be set up, which can only be done after the transfer to the previous receiver is complete and the circuit to that receiver is released.
  - Multicast: can save bandwidth and computing resource

New options: multicast and P2P

- Multicast:
  - Hypothesis: total delay for distributing the data to the receivers will be lower for a given computing capacity of the upstream servers, or conversely, the same transfer delay can be achieved as with IP-routed service but with smaller upstream server computing capacity.
  - Negative: one or more slow receivers can slow down everyone
- P2P: under consideration by NCAR
  - Hypothesis: transfer delay does not increase with number of receivers as with unicast TCP, but per product, but 90% transfer delay will be more than with multicast

Multicast VCs

- Difference between multicast VC and multicast IP routed service
  - No congestion related losses on multicast VC
  - In-sequence delivery of packets
- Multicast Transport Protocols:
  - For IP-routed service: RMTP, SRM, MTP-2, Ramp, etc.
  - No multicast transport protocol for virtual circuits
    - To our best knowledge
  - Hence designing a reliable Multicast Virtual Circuit Transport Protocol (MVCTP)

19

Why change status quo?

- Subscribers: The last of the 104 subscribers will see higher latency in receiving the new data product; but this is not significant
- NCAR: maintain 9 servers + high access link rate
- Others: are these "continuous" LDM flow unfair to others? Jain fairness:

$$f(x) = \frac{(\sum_{i=1}^{n} x_i)^2}{n \cdot \sum_{i=1}^{n} x_i^2} \quad x_i \geq 0$$

20

Planned Work

- Prototyping and evaluation of MVCTP
- Modification of LDM for execution over MVCTP (MVC-LDM)
- Design of a hybrid solution to use MVCs in the core network and unicast TCP connections in edge networks (Hybrid-LDM)
- Analytical model for multicast
- Fairness and impact on other flows study: OWAMP delays + simulations

Agenda check

- **EAGER Project Review**
  - 9:30: EAGER introduction (MV)
  - 9:40: EAGER: IDD data characteristics (Matt Manley)
  - 9:55: EAGER: Year 1 report (MV)
  - 10:15: EAGER: Multicast transport protocol for VCs (Jie Li)
- **SDCI Project Review**
  - 10:30: SDCI year 1 UVA and workplan for UVA and UNH (MV)
  - 11:00: SDCI year 1 UCAR workplan: John Dennis, NCAR
  - 11:30: GridFTP, Ganglia, ANI testbed: Zhengyang Liu
  - 11:45: GridFTP data analysis: J. Gradon Koelling and Peter Sahajian
- 12-1: Lunch with discussion

SDCI Project objectives

- Data analysis
  - CESM wide-area network usage
  - CESM intra-datacenter network usage
- Develop integrated networking solutions
  - RoCE intra-datacenter with wide-area VC
  - iWARP intra-datacenter with wide-area IP
- Technology transfer to CESM and other scientific communities

23

First steps

- Since start of project (Sept. 15)
  - Recruited students
  - Students are learning background material
  - PIs Year 1 planning weekly meetings (3)
  - Project and Collab Web sites set up
  - Setting up computer accounts
    - DOE ANI testbed (Liu will present this information)
    - Magellan NERSC (compute cycles + network node access)
    - Magellan ANL (approved by Rick Stevens)

24

Year 1 overall work plan

- Data analysis to gain an understanding of CESM projects' wide-area networking usage: UVA and NCAR
- Intra-datacenter MPI applications: evaluate IB, RoCE and iWARP interconnects for a subset of CESM applications, and other benchmarks: NCAR, UNH and UVA (analysis)
- Software implementation
  - EXS library: UNH
  - Integrate IDCIM with file transfer applications: UVA
  - Use NMI Build and Test service
- Broader impact activities: all
  - Prepare course module on datacenter networking
  - Participate in diversity related activities

25

UNH Year 1 Workplan

- Complete EXS library development and documentation
- Intra-datacenter MPI applications
- Course module on data networking
- Diversity activities

UNH EXS introduction

- Earlier papers show good throughput performance for a proof-of-concept implementation of Extended Sockets (EXS) as a user-level interface to Remote Direct Memory Access (RDMA) technologies.
- Since then we have prepared a series of simple programs that utilize the Open Fabrics Alliance (OFA) verbs to exercise the various features of the three current RDMA technologies:
  - InfiniBand (IB)
  - RDMA over Converged Eehernet (RoCE)
  - iWARP (internet Wide Area RDMA Protocol)

UNH EXS workplan

- Oct-Dec 2011 quarter:
  - use these programs to collect, compare and evaluate measurements on the performance of the three RDMA technologies.
- Jan-Mar and Apr-Jun 2012 quarters:
  - utilize these results to:
    - improve the performance of EXS
    - enhance the EXS implementation to enable it to dynamically adapt to the differences in the RDMA technology being employed, since some features available in IB and RoCE are not available in iWARP
    - extend the EXS API to allow user programs to tune EXP performance to match their application. For example, to trade off latency for cpu time.
- Jul-Sep 2012 quarter:
  - complete the implementation and documentation of EXS and construct a series of tests to demonstrate its performance.

UNH workplan: Intra-datacenter MPI applications

- Work with NCAR to execute applications and collect measurements to evaluate IB, RoCE and iWARP on intra-datacenter MPI applications (to be selected).
- Work with NCAR and UVA to write reports and publish papers on these results.
- Timeline to be agreed upon jointly, and is conditional on availablity of equipment at the national labs.
- First step is to select appropriate benchmarks and measurement tools.
  - Need to evaluate their current implementations on RDMA technologies, the feasibility of converting them if necessary, and the availability of testbeds on which to run them.
- Next step is to set up and run the benchmarks and measurement tools to gather statistics for analysis.

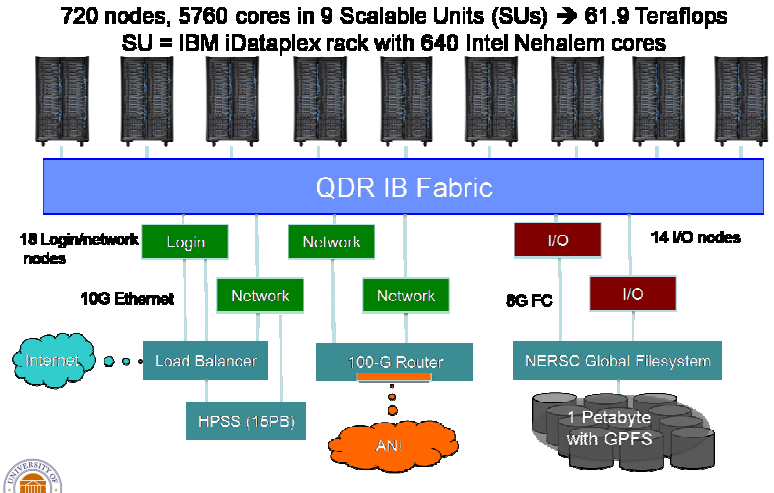
UNH workplan: Broader impact

- Work with UVA and NCAR to develop a course module on datacenter networking.
  This involves several parts:
  1. Basic concepts
  2. Intra-datacenter technologies (InfiniBand)
  3. Inter-datacenter technologies (Virtual Circuits, iWARP)
  4. Remote Direct Memory Access (RDMA)
  5. Converged ethernet - motivation, current state, RoCE
  6. MPI applications
  7. Future directions
- Participate in diversity related activities

UVA data analysis

- Determine the reasons for poor end-to-end application-level performance
  - Wide-area file transfers
    - Use same strategy as developed in EAGER
      - Characterize data product arrival process
      - Characterize data product sizes
      - Characterize network-wide view of sender-receiver nodes
      - Determine which network service type best suits CESM's project needs
    - Type of data to obtain:
      - Liu, Koelling and Sahajian presentations
  - Intra-datacenter MPI applications
    - NCAR and UNH will obtain measurements by executing CESM applications and benchmarks on Magellan compue cluster; CESM apps need 1000 cores
    - UNH will obtain low-level measurements on Magellan network nodes using RoCE and iWARP
    - UVA will combine in analytical models

Magellan @ NERSC

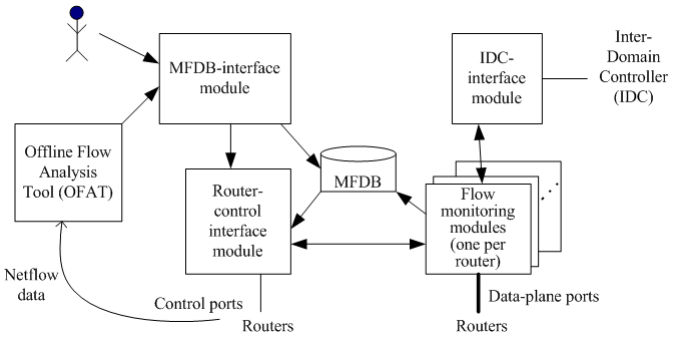


UVA software development

- Integrate IDCIM with file transfer applications
- IDCIM: InterDomain Controller Interface Module: VC scheduler
- CESM scientists use GridFTP
- Raj Kettimuthu, ANL, said they are working on integrating GridFTP and OSCARS
- So we plan to use scripts and not change GridFTP code

33

Hybrid Network Traffic Engineering Software (HNTES)



- MFDB: Monitored Flow Data Base
- Some components can be centralized and rest distributed

DOE project: Leverage this work

34

Components contd.

- IDC Interface Module (IDCIM)
  - interfaces with Inter-Domain Controller (IDC)
- IDC (not part of HNTES software)
  - VC scheduler
  - Accepts reservations (starting time: now allowed)
  - Provisions circuits at scheduled start time
  - Releases circuits
  - Sets Policy Based Route in IP router to redirect packets from default IP-routed path to newly established circuit
  - Removes PBR entry when circuit is released
- DOE ANI testbed has a running IDC
  - Liu will describe this testbed

35

IDCIM demo: Oct. 2010 (DOE)

- IDC Interface Module (IDCIM) run on host zelda2 at UVA
- Started with OSCARS Java Client
- Connect to OSCARS test IDC server and subscribe for notifications
- Automatic signaling

36

IDCIM demo

- Run client
- Send request for circuit from client to IDCIM
- IDCIM (Java) packages per IDCP and sends to IDC with "now" request and automatic signaling
- When PathSetup is confirmed, IDCIM notifies client (FMM)
- IDCIM demo video: UVA's DOE project site:
  - http://www.ece.virginia.edu/mv/research/DOE09/talks/annual-review-oct10/IDCIM.wmv

37

Agenda

- **EAGER Project Review**
  - 9:30: EAGER introduction (MV)
  - 9:40: EAGER: IDD data characteristics (Matt Manley)
  - 9:55: EAGER: Year 1 report (MV)
  - 10:15: EAGER: Multicast transport protocol for VCs (Jie Li)
- **SDCI Project Review**
  - 10:30: SDCI year 1 UVA and workplan for UVA and UNH (MV)
  - 11:00: SDCI year 1 UCAR workplan: John Dennis, NCAR
  - 11:30: GridFTP, Ganglia, ANI testbed: Zhengyang Liu
  - 11:45: GridFTP data analysis: J. Gradon Koelling and Peter Sahajian
- 12-1: Lunch with discussion

38